A Measure of Skill in Forecasting a Continuous Variable

IRVING I. GRINGORTEN
Air Force Cambridge Research Laboratories, Bedford, Mass.

(Manuscript received 27 June 1964)

ABSTRACT

To test the skill of a forecaster the rule for the score S, for the quantitative forecast of temperature or a similar variable, becomes $S=-\ln(1-P_1)P_2-1$ where P_1 is the cumulative climatic frequency of the forecast value T_F, or the cumulative climatic frequency of the subsequently verified value T_V, whichever is smaller. The value P_2 is the greater of these two frequencies. Such frequencies must be made conditional to the initial state of the weather in order to properly reward forecasters for recognizing future changes in the weather. For the quantitative forecast of precipitation, or similar variables, there are several alternate formulas for skill scores, each formula depending upon whether or not any precipitation is forecast or observed, or both forecast and observed.

This system of scoring assures that unskilled strategies, such as the forecasting of the most frequent values, or persistence forecasting, will net the forecaster an expected average of zero. For individual accurate forecasts the rewards are greatest but still depend on the frequency or infrequency of the verified events. For inaccurate forecasts the rewards can be positive or negative, depending upon the sign and amount of change that is predicted, as well as the subsequent verification.

1. Introduction

A system of scoring forecasts of a continuous variate, such as a forecast, 24 hours in advance, of surface temperature, or the maximum daily temperature or the 24-hour precipitation amount, has been devised to test the skill of the forecaster. It encourages the forecaster to give the best estimate of the quantity, without a desire to hedge toward a normal figure, or without an unwarranted desire to try for the spectacular forecast. While there have been many previous papers written on the scoring of forecasts (Heidke, 1926; Brier and Allen, 1951; Gringorten, 1951; Vernon, 1953) they have been directed either toward the test of the forecaster's ability to select a category of events or they have judged the forecasts by the absolute error or square of the error of the forecasts (Lorenz, 1959). The latter is essentially a test of the accuracy of the forecast instead of the skill of the forecaster. Probability estimates of subsequent events have been limited by previous verification programs to the estimation of probabilities of mutually exclusive categories (Sanders, 1963; Root, 1962).

Forecasting skill, in this paper, is defined as follows: the ability of the forecaster to sort or classify the existing or previous state of the weather so that, within his classification, the probability of one subsequent event is increased above its conditional climatic frequency, conditional to the existing value or category of the weather that is being predicted.

Beginning with this definition the scoring system should be composed to eliminate any advantage to an unskilled strategy, such as an uninformed forecast of "no rain," the most frequent amount, or a continuous declaration that the existing weather will persist. If probability statements are requested, there should be no strategy that will lessen the penalty for uncertainty, such as quoting the basic climatic frequencies of the future events.

2. Derivation of scores

For mutually exclusive categories of the weather. In the 1951 paper by this author it was shown that the above conditions can be met in the forecasting of a set of n mutually exclusive events $X_1, X_2, \cdots X_n$ by composing a set of scores, such that the score for a correct forecast is inversely proportional to the conditional climatic frequency ${}_cP_i$ of the correctly forecast event. The score for error is zero. For each initial condition a table of scores would resemble Table 1.

At the time of publication of the 1951 paper, it was expected that the forecaster would make a selection of a single event as his forecast. It is possible, however, to test the forecaster's skill in a set of probability estimates, say, of three mutually exclusive categories:

$$ {}_fP_1,\ {}_fP_2,\ {}_fP_3, $$

where the prescript f denotes a forecast probability. In view of the above definition of skill, it is desirable to test whether each of the three probabilities have been

TABLE 1. Example of a set of scores for each combination of forecast and verification of three mutually exclusive events (X_1, X_2, X_3) whose climatic frequencies are, respectively, 50%, 30%, 20%.

Forecast	Verified X_1	X_2	X_3
X_1	2.0	0	0
X_2	0	3.3	0
X_3	0	0	5.0

properly increased or decreased, above or below, the climatic frequency. That is, we should test the ratios

$$ {}_fP_1/{}_cP_1, \quad {}_fP_2/{}_cP_2, \quad {}_fP_3/{}_cP_3. $$

The probability figures can be interpreted as positive or negative statements on the subsequent events. If

$$ {}_fP_1/{}_cP_1 > 1.0 $$

then the eventual occurrence of the first category of weather would be considered a positive verification, for which the forecaster would receive a score of $1/{}_cP_1$. If

$$ {}_fP_2/{}_cP_2 > 1.0 $$

then the nonoccurrence subsequently of the second category would be considered an unsuccessful verification, for which the forecaster would receive a score of zero. If

$$ {}_fP_3/{}_cP_3 < 1.0 $$

the nonoccurrence subsequently of the third category would be considered a successful verification, for which the score would be $1/(1-{}_cP_3)$. This system would treat each event dichotomously: it either occurs or does not occur.

After a sizable sample of forecasts have been collected, the scores, added together, will have an expected average of 1.0 for "no skill" and an expected average of 2.0 for constantly perfect forecasts. "No skill" forecasts include such types as a perpetual forecast of a clear sky, or a perpetual forecast of an overcast sky, or a forecast that the present weather will persist, or a selection of category by random process, such as picking a number out of a hat. The measure of skill, on a scale from zero to one, is obtained by subtracting unity from the average

$$ \sum_{i=1}^{n} \sum_{j=1}^{N} \beta_{ij}/nN, $$

where n is the number of mutually exclusive categories, N is the number of days on which the probability forecasts are made, and β_{ij} is the score awarded for the ith category on the jth day ($=1/{}_cP_i$, or zero, or $1/(1-{}_cP_i)$).

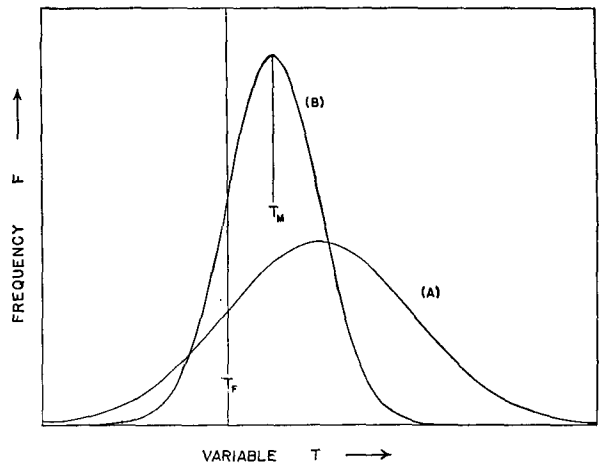

FIG. 1. Climatic frequency of a variable T is represented by curve (A). A probability forecast of T is represented by curve (B), showing a concentration on one value (T_M) with relatively little deviation.

Probability forecast of a continuous variable. The above system is provided as a preliminary requisite to the development of a system for the continuous variable. In Fig. 1 curve (A) represents the climatic frequency distribution of the variate T, say temperature, assumed given for an initial state. Curve (B) represents the probability forecast statement, which, in Fig. 1, implies that the forecaster is centering his probabilities on one modal value T_m with relatively little deviation. There is a temperature T_F at which the forecaster's cumulative probability is equal to the climatic cumulative frequency. For values of $T<T_F$ the forecaster's cumulative probability P_F is smaller than the climatic frequency P. This is equivalent to a negative statement on a temperature equal to or less than T. If the verified temperature T_V is greater than T, then the prediction curve is considered as a successful statement with respect to T, for which the score is $1/(1-P)$. But, if the verified temperature T_V is equal to or less than T, then the forecast is unsuccessful with respect to T, for which the score is zero.

For values of $T>T_F$ the forecaster's indicated probability of a temperature of T, or less, is greater than the climatic frequency P. If, then, the verified temperature T_V is greater than T, then the forecast statement is considered unsuccessful with respect to T, for which the score is zero. But, if the verified temperature T_V is less than T then the forecast with respect to T is considered to earn the score $1/P$.

This leads to the following result:

If $T_V>T_F$ then the score S', averaged with respect to all possible values of T, is

$$ S' = \int_0^{P_F} \frac{dp}{(1-p)} + \int_{P_V}^{1} \frac{dp}{p} = -\ln(1-P_F)P_V. \quad (1) $$

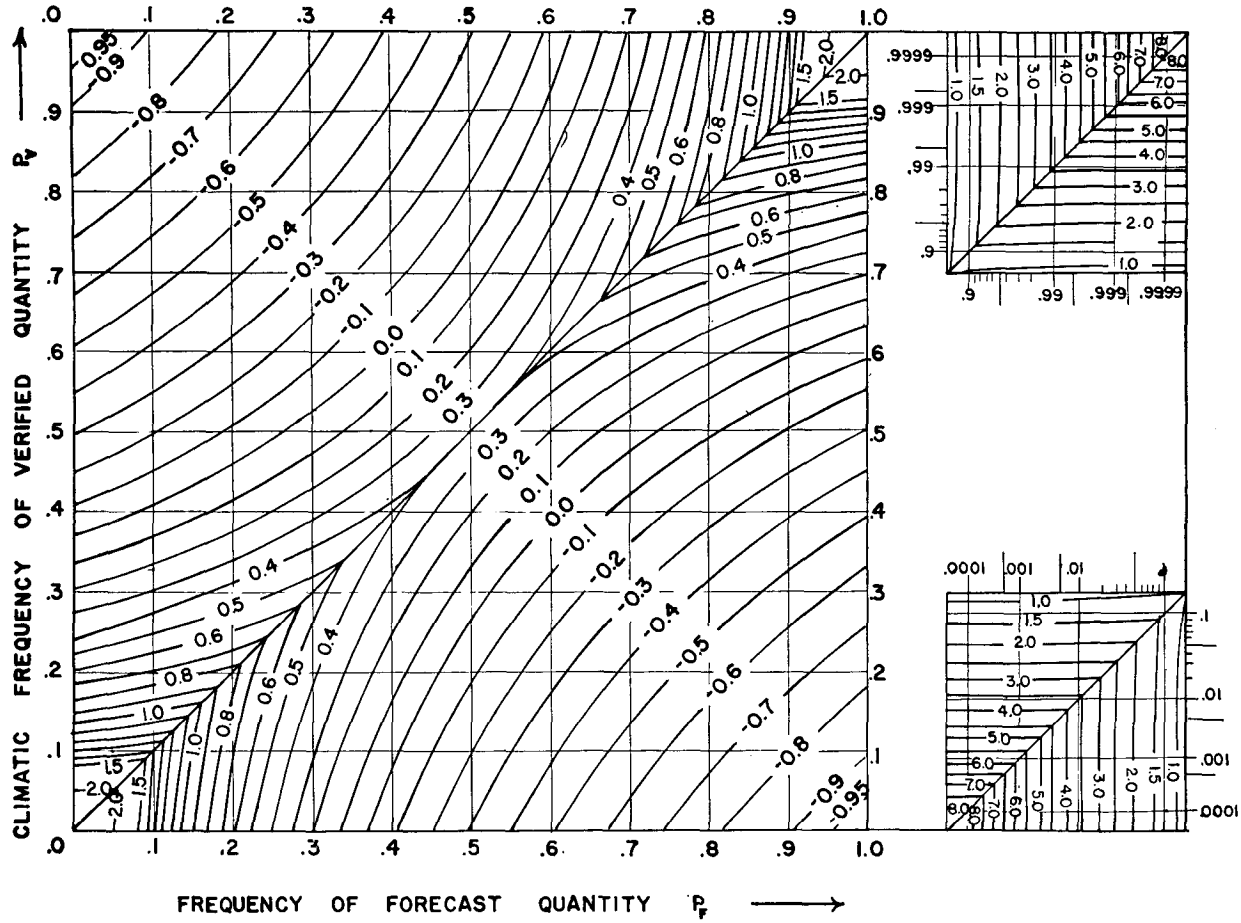

FIG. 2. A nomogram for the computation of individual scores. For each forecast T_F and the subsequent verification T_V the cumulative climatic frequencies P_F and P_V are entered in the chart. Their cross section gives the earned score. For both values P_F and P_V less than 0.1 or greater than 0.9 the inserted charts are provided for better resolution.

If $T_V < T_F$, then

$$S' = \int_0^{P_V} \frac{dp}{(1-p)} + \int_{P_F}^{1} \frac{dp}{p} = -\ln(1-P_V)P_F. \quad (2)$$

If $T_V = T_F$, or $P_V = P_F$, then the score is the maximum possible.

For the verification:

$$S' = \int_0^{P_V} \frac{dp}{(1-p)} + \int_{P_V}^{1} \frac{dp}{p} = -\ln(1-P_V)P_V. \quad (3)$$

Selection of a single value of the continuous variable. A single forecast value becomes identical with T_F of the previous section. Hence, the scores (1,2,3) apply, respectively, to the verification T_V greater than, less than, or equal to, T_F. The expected average score for no skill is 1.0. For perfect forecasting, on the average, the expected value is 2.0. Hence, subtracting unity from the value of $-1n(1-P_F)P_V$ or $-1n(1-P_V)P_F$ will provide an indication to the forecaster of gain or loss in the score total. Individual scores can range from -1.0 to infinity. But the average score for unskilled forecasts will be zero, and for perfect forecasts will be 1.0.

A nomogram for the computation of the score, S, is given in Fig. 2.

$$\begin{aligned} S &= -1n(1-P_F)P_V - 1 \quad \text{when} \quad P_F < P_V \\ &= -1n(1-P_V)P_F - 1 \quad \text{when} \quad P_V < P_F \\ &= -1n(1-P_F)P_F - 1 \quad \text{when} \quad P_V = P_F. \end{aligned} \quad (4)$$

Fig. 2, to find S, becomes the working tool of the scoring system.

Prediction of rainfall amounts. Ordinarily there is a large probability of no measurable precipitation at a station on any given day. The climatic frequency curve, therefore, will resemble Fig. 3. Where R_F is the predicted rainfall amount, P_F the cumulative climatic frequency of R_F, R_V the verified amount and P_V its cumulative climatic frequency, if

$$R_F = R_V = 0 \quad \text{or} \quad P_F = P_V = P_0,$$

where P_0 is the climatic frequency of no rain, then the score, averaged with respect to all positive amounts is

$$S'=\int_0^{P_0}\frac{dp}{P_0}+\int_{P_0}^{1}\frac{dp}{p}=1-\ln P_0. \quad (5)$$

If $R_F=0$, $R_V>0$, then

$$S'=\int_{PV}^{1}\frac{dp}{p}=-\ln P_V. \quad (6)$$

If $R_F>0$, $R_V=0$, then

$$S'=\int_{PF}^{1}\frac{dp}{p}=-\ln P_F. \quad (7)$$

If both forecast and verification show measurable precipitation, then, if $R_F>R_V>0$,

$$S'=\int_0^{P_0}\frac{dp}{1-P_0}+\int_{P_0}^{PV}\frac{dp}{1-p}+\int_{PF}^{1}\frac{dp}{p}$$
$$=P_0/(1-P_0)+\ln(1-P_0)-\ln(1-P_V)P_F. \quad (8)$$

If $R_V>R_F>0$,

$$S'=P_0/(1-P_0)+\ln(1-P_0)-\ln(1-P_F)P_V. \quad (9)$$

These results are summarized in Table 2, with S' corrected to $S=(S'-1)$. The expected average of S ranges from zero to one. Table 2 will apply to any other variable that has a high frequency of the zero value such as the calm condition of wind speed.

3. Method illustrated

Quantitative forecast of temperature. The example selected is the January temperature at 1200C at Minneapolis, Minn. Fig. 4 shows the temperature frequencies in degrees Fahrenheit for three Januaries, 1943, 1944, 1945. The conditional distributions are classified according to the previous day's temperature. The data for the short record of 93 days show an anticipated instability. Nevertheless, the climatology as depicted in Fig. 4 should be used, insofar as possible, for the verification and scoring of 24-hour forecasts in the three months of the record. The solid parts of the curves in Fig. 4 have been obtained by interpolation, the dashed parts by extrapolation. The latter procedure

TABLE 2. Skill scores, in terms of the climatic frequency P_0 of no daily precipitation, the cumulative climatic frequency P_F of the predicted amount and P_V of the verified amount. It is assumed that the frequencies P_0, P_F, P_V are made conditional to the initial state of the weather.

Forecast	Verify No rain or trace	Verify Measurable amount
No rain or trace	$-\ln P_0$	$-\ln P_V-1$
Measurable amount	$-\ln P_F-1$	$\frac{*P_0}{1-P_0}+\ln(1-P_0)-\ln(1-P_1)P_2-1$

* P_1 is the cumulative climatic frequency P_F or P_V, whichever is smaller. P_2 is the greater frequency.

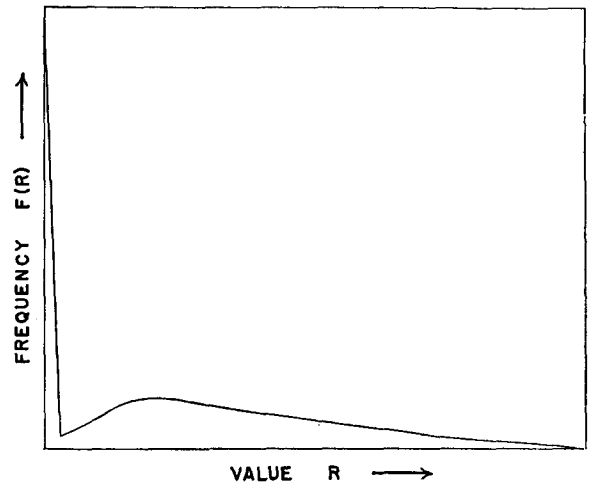

FIG. 3. Climatic frequency of a variable such as daily rainfall, that has a high frequency of the value zero.

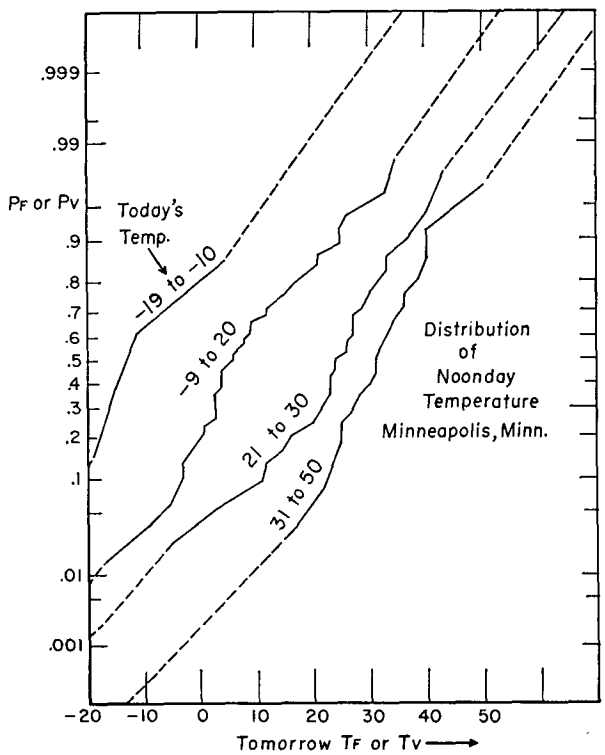

FIG. 4. The cumulative climatic frequency of noonday temperature at Minneapolis, Minn., in the three Januaries of 1943, 1944, and 1945. The frequencies are conditional to the previous day's noonday temperature.

TABLE 3. A sample of forecasts T_F, observed values T_V and automatically the previous day's values, the cumulative climatic frequencies P_F, P_V of forecast and observed values, and finally the score S for each day's forecast. The example is for noonday temperature at Minneapolis, Minn.

Date	Forecast T_F	Forecast P_F	Observed T_V	Observed P_V	Score S
1 Jan 1943	—	—	26	—	—
2	15	0.17	15	0.17	+1.0
3	10	0.68	16	0.77	+0.4
4	20	0.85	0	0.20	−0.6
5	15	0.76	3	0.36	−0.3
6	8	0.59	8	0.59	+0.4
7	0	0.20	15	0.76	−0.5
8	0	0.20	9	0.66	−0.4
9	35	0.985	35	0.985	+4.0
10	25	0.22	24	0.136	+0.7

TABLE 4. The average skill scores by various methods, skilled and unskilled, on the 93 days in January 1943, 1944, 1945 at Minneapolis, Minn. The 24-hour forecasts were of temperature, verifying at 1200C.

Description of forecast	Ave. score	Expect
A constant forecast of −16F	−0.01	0.00
A constant forecast of 0F	0.00	0.00
A constant forecast of 16F (the climatic mean)	−0.03	0.00
A constant forecast of 32F	−0.02	0.00
24-hour persistence	0.08	0.00
Random numbers between −20 and 50	0.02	0.00
Forecast is −20F, −10F, 0F, 10F, 20F, 30F or 40F, always within 5F of verification	0.75	
Perfect (to the nearest degree)	0.91	1.00

TABLE 5. Sample scores for various combinations of initial temperature, forecast T_F and verification T_V for January noonday temperatures at Minneapolis, Minn.

Initial	Forecast T_F	Verification and scores −16F	0F	16F	32F
−10F	−16F	0.6	−0.4	−0.6	−0.6
30F	−16F	4.5	2.1	0.6	−0.8
−10F	0	−0.4	0.8	0.6	0.6
30F	0	2.2	2.2	0.6	−0.7
−10F	16F	−0.6	0.6	3.2	3.1
30F	16F	0.6	0.7	0.8	−0.6
−10F	32F	−0.6	0.6	3.0	7.1
30F	32F	−0.8	−0.8	−0.6	0.9

is admittedly subjective, but the evidence below indicates relatively little effect on the program.

In Table 3, the method of scoring is illustrated. For each predicted temperature T_F the cumulative climatic frequency P_F was found in Fig. 4. Likewise, for each verification, T_V, the cumulative climatic frequency P_V was found. From Fig. 2 the score S was obtained for each pair (P_F,P_V).

The results for 93 days, consisting of the average score for each of several methods of forecasting, both skilled and unskilled, are shown in Table 4. For nonskilled methods of forecasting the average, as anticipated, is close to zero. For perfect forecasting the average is close to 1.0. The forecaster who is able to predict the 24-hour change of temperature, always within 5 Fahrenheit degrees, earns a score that is 75% of the expected perfect score.

Table 5 has been prepared to indicate how the scores vary with the magnitude of the predicted change and with the rarity of the verified event, as opposed to the numerical difference between forecast and verified values. For the same combination of forecast and verification it is possible that the forecaster either will be rewarded for skillful forecasting or be penalized, depending upon the antecedent condition.

Quantitative forecast of precipitation. The example selected is for 24-hour precipitation, Dec. 1963, Jan. 1964, Feb. 1964 at Logan International Airport, Boston, Mass. The forecasts were made by the U. S. Weather Bureau, of the amount of precipitation to fall between 0700E of each day to 0700E of the following day. The deadline time for the forecast was 0100E. Classified according to the previous 24-hour amount up to 0100E, the 91 values of subsequent precipitation yielded the cumulative frequency distributions of Fig. 5. (The solid parts of the curves were drawn by interpolation, the dotted parts by subjective extrapolation.) The scores for the various combinations of initial condition, forecast and verification are shown in Table 6. Values of $-\ln P_2-1$ are obtainable from the horizontal axis of Fig. 2. The results for several types of forecasts, skilled and unskilled, are shown in Table 7.

Assuming that a verification program is to be conducted, there would be an advantage in having a set of climatic tables or charts in the forecasting office. From such charts and Fig. 2 of this paper, a forecaster would obtain his approximate score for each day's forecast and would be able to find tentative running averages. To test this usability, the conditional climatic distributions of daily precipitation in the five winter months, 1958-63 were determined and plotted (Fig. 6). From these the climatic frequencies P_V of the values observed in the 1963-64 winter season were obtained, as well as the climatic frequencies P_F of the U. S. Weather Bureau's forecast values. For the 91 days in the 1963-64 season the "perfect" average was 1.07 instead of 0.97 (Table 7). The forecasters' average was 65% instead of 49%. If appears, therefore, that a running average could be obtained tentatively until a more representative climatology is available for ultimate scoring.

4. Remarks

There are several advantages to the system of scoring, described above, that make it superior to previous systems. It is a measure of skill, not of accuracy. The individual scores measure the significance of the forecast relative to the amount of change predicted, the

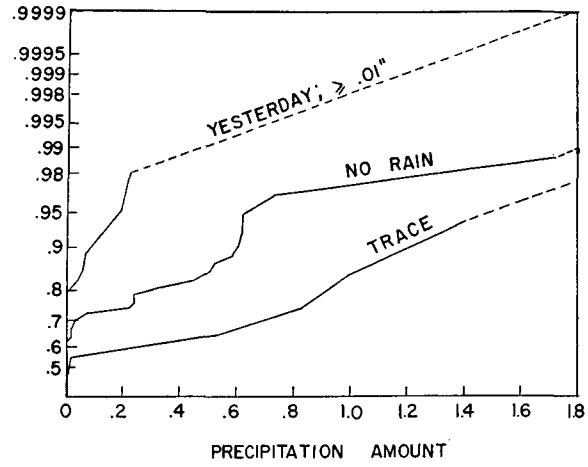

FIG. 5. Cumulative frequency distributions of daily precipitation amounts during the winter of Dec. 1963, Jan., Feb., 1964 at Boston, Mass. airport. The three distributions are classified by the previous day's precipitation.

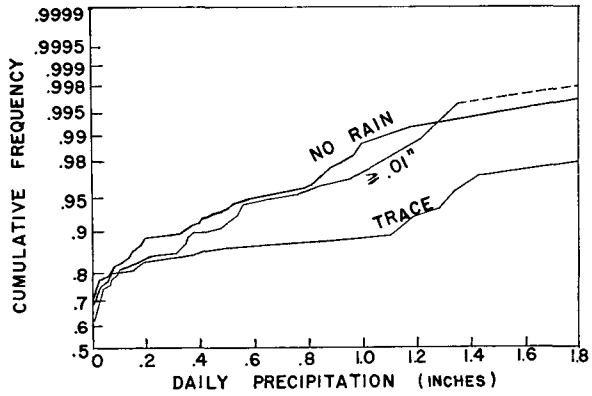

FIG. 6. Cumulative frequency distributions of daily precipitation amounts during the winters of Dec. 1958 to Feb. 1963 at Boston, Mass., airport.

sign of the change as well as the rarity of the verified event. Forecasters will earn positive scores that are fractions of the scores for perfect forecasting. Missing a rare event does not necessarily become a tantalizing experience to the forecaster, because he can earn a relatively high score by predicting correctly the sign of the change in the weather. Yet there is no substitute, and no compromise, for a wholly accurate forecast. There is no rule of thumb that will serve the forecaster if he is uncertain about the weather. The system, therefore, forces the forecaster to pursue the analysis of the situation until the probabilities are sharpened around at least one value.

Forecasting for the test of skill is compatible with forecasting for operational utility. Probability statements that are prepared for an operations office can be examined to determine at what value the forecast cumulative probability is equal to the climatic frequency. This value becomes the value to be tested in

TABLE 6. The scores that are awarded to the forecaster for each combination of amount of precipitation of previous day, amount forecast and amount verified, at Boston, Mass., Dec. 1963, Jan., Feb. 1964.

Initially on day of forecast	Forecast	Verification and score None or trace	≥0.01
None	None or trace	0.5	$-\ln P_V-1$
Trace	None or trace	0.8	$-\ln P_V-1$
≥0.01 inches	None or trace	0.2	$-\ln P_V-1$
None	≥0.01 inches	$-\ln P_F-1$	$S_{FV}+0.7$*
Trace	≥0.01 inches	$-\ln P_F-1$	$S_{FV}+0.2$*
≥0.01 inches	≥0.01 inches	$-\ln P_F-1$	$S_{FV}+2.2$*

* $S_{FV}=-\ln(1-P_1)P_2-1$ where P_1 is the cumulative climatic frequency P_F or P_V, whichever is smaller. P_2 is the greater frequency.

TABLE 7. The average skill score for 91 days of forecasting, by several methods, skilled and unskilled, of 24-hour precipitation at Boston, Mass, Airport Dec. 1963, Jan., Feb. 1964.

Method	Ave. score	Expected score
Rule of thumb*	0.01	0.00
Forecast by chance (random numbers)	0.01	0.00
U. S. Weather Bureau forecasts	0.49	
Perfect (to nearest .01 inches)	0.97	1.00

* The rule was to forecast no precipitation if previous amount was measurable or zero; but the amount of 0.08 inches was predicted if previously there was a trace. (Frequency of precipitation following trace is 0.55.)

the verification program. This paper, however, has not attempted to present a scoring system for forecasts that present two or more alternative modal values with associated probabilities. Scoring of such forecasts can be done, and might be presented in a follow-up paper to this one.

If a system of scoring, such as described in this paper, were to be initiated in a weather office, it would prompt the forecasters into preparing climatological tables or charts of frequency distributions of temperature, dew point, wind speed, precipitation amount, and so on. A running average of tentative scores could be kept. But the ultimate scores would depend upon the specialized climatology of those days for which forecasts would be made. In this way, the "perfect" average would be close to 100% and the unskilled techniques, including persistence forecasting, would earn the expected score of zero.

The value of the scoring system in the verification of prognostic charts is apparent. There need not be a serious penalty for a large mean square error if, say, a system deepens much more than expected. On the other hand, the scores will be greatest for an exact prognosis.

The suggested program should end the long search that began during World War II (U. S. Army Air Forces, 1943) for a satisfactory system of verification

of continuous variables or amounts. Or will it be just the beginning of a renewal of the search?

Acknowledgment. The U. S. Weather Bureau forecasts and the observed amounts were graciously provided, for this study, by Dr. O. Tenenbaum, meteorologist-in-charge, Boston office.

REFERENCES

Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. *Comp. Meteor.*, Boston, Amer. Meteor. Soc., 841–848.

Gringorten, I. I., 1951: The verification and scoring of weather forecasts. *J. Amer. Stat. Assoc.*, **46**, 279–296.

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen in Sturmwarnungsdienst. *Geografike Annaler*, **8**, 310–49.

Lorenz, E. N., 1959: *Prospects for Statistical Weather Forecasting.* Mass. Inst. Tech., Dept. Meteor., Statistical Forecasting Project, AF 19(604)–1566, Final Report.

Root, Halbert E., 1962: Probability statements in weather forecasting. *J. Appl. Meteor.*, **1**, 163–168.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.

U. S. Army Air Forces, Weather Information Branch, 1943: *Short Range Forecast Verification Program*, Report No. 602, Washington, D. C.

Vernon, E. M., 1953: A new concept of skill score for rating quantitative forecasts. *Mon. Wea. Rev.*, **81**, 326–329.